

Semantic MediaWiki в действии: опыт построения семантического портала

Даниэль М. Херциг and Базил Эль

Институт AIFB
Технологический Институт Карлсруэ
76128 Карлсруэ, Германия
{herzig, basil.ell}@kit.edu
<http://www.aifb.kit.edu>

Translated into Russian by Yury V. Katkov,
katkov.juriy@gmail.com

Аннотация Вики-системы позволяют пользователям совместно создавать и управлять содержимым веб-сайта. Семантические вики предоставляют дополнительные средства для семантической аннотации контента, тем самым улучшая структуру веб-сайта. В данный момент семантические вики переживают подъем популярности, ведь преимущество структурированных данных перед неструктурированными очевидно. В статье приведено описание опыта создания портала института AIFB на базе движка семантической вики Semantic MediaWiki. Этот опыт должен проиллюстрировать то, как получить максимум от использования семантических вики для создания семантических порталов.

В статье рассмотрены вопросы проектирования портала, в частности то, что можно эффективно совместить ввод с помощью форм со свободным семантическим вики-аннотированием. Мы также используем заранее предопределенную схему данных, что позволяет сделать представление знаний на портале гибким, расширяемым и структурированным. Далее обсуждаются вопросы, касающиеся того, как эти структурированные данные меняются со временем и то, как зависит их гибкость в зависимости от изменений. Эти вопросы иллюстрируются статистикой, основанной на реальных данных, находящихся на портале. В продолжение представлены особенности использования структурированных данных и то, какие преимущества это дает. Так как все преимущества имеют свою цену, мы также провели работу по исследованию производительности Semantic MediaWiki и сравнили результаты с производительностью не семантической платформы MediaWiki.

Наконец, мы показываем как существующие техники кэширования могут быть применены для повышения производительности.

1 Введение

Веб-порталы являются точками входа для тех, кто заинтересован в получении или обмене информацией, посвященной какой-либо организации или

теме. Как правило, веб портал поддерживается сообществом заинтересованных пользователей. Использование семантических технологий и семантически размеченного контента для порталов доказало свою полезность довольно давно (см. например [1]), однако вопросы, связанные с предоставлением структурированных данных привлекли много внимания несколько позже благодаря инициативе Открытых Связанных Данных (Linked Open Data). Вместе с тем заметим, что упомянутые в citemaedche2003sp-tsa подходы делали акцент на использовании формальных онтологий, создающихся усилиями инженера по знаниям до того, как будут реализованы конкретные приложения. Результатом работы инженеров по знаниям является формальная, согласованная и выразительная структура априорных знаний [1,2]. Формирование подобной структуры - это довольно трудоемкий процесс, и он требует дополнительных усилий в случае, если эта структура постоянно меняется и корректируется. Помимо этого недостатка, в работе [3] отмечается, что системы контроля версий структурированной информации находятся сейчас в зачаточном состоянии. Столь важные при совместной работе средства взаимодействия между участниками сообщества также не удовлетворяют современным требованиям. Недавно проведенное исследование [4] показало как популярная CMS (система управления содержимым, content management system) *Drupal*, которая будет поддерживать семантические данные начиная с версии 7, может быть применена для построения семантических приложений. Мы развиваем другой подход, привлекая к созданию и управлению данными сообщество пользователей.

Вики является одной из наиболее успешных технологий для поддержки сообществ по интересам во Всемирной Паутине. Вики-системы используют для совместного создания и управления содержимым в текстовом и полуструктурированном виде. Главная идея, лежащая в основе вики состоит в том, чтобы воодушевлять людей вносить свой вклад в проект, предоставляя настолько простые в освоении средства, насколько это возможно. Такой подход к созданию контента мы называем community-driven - управляемый сообществом. При этом именно сообщество пользователей осуществляет контроль за процессом разработки содержимого сайта. Семантические вики позволяют аннотировать содержимое для того, чтобы снабдить его структурой, что позволяет рассматривать вики как частично структурированную базу данных. Также становится возможным производить запросы к структурированному содержимому. Запросы позволяют использовать данные, хранящиеся в вики, и создавать различные отображения этих данных, благодаря чему вики становятся еще более мощными системами управления контентом. Более того, благодаря семантическим аннотациям, структурированные вики-данные становятся доступными для внешних сервисов, в том числе для сервисов Linked Open Data.

В этой работе мы описываем использование семантической вики *Semantic Media Wiki*¹ [5,6] (SMW) для создания портала нашего института, который доступен по адресу <http://www.aifb.kit.edu>. Портал служит представи-

¹ <http://semantic-mediawiki.org/>

тельством института AIFB, академического учреждения с размером штата около ста пятидесяти человек. Портал является полноценным Semantic Web приложением и содержит примерно 16 тысяч страниц, на которых расположены около 105 тысяч семантических аннотаций. В таблице 1 приведены численные характеристики портала.

Вики-системы обычно предоставляют возможность свободного аннотирования контента семантическими аннотациями, однако существуют средства, позволяющие ограничить используемый при аннотировании словарь терминов. Это реализуется с помощью ввода, основанного на формах. В работе [2] подчеркивается важность соблюдения баланса между использованием неструктурированного содержимого (что, все-таки лучше, чем полное его отсутствие) и работой со структурированными данными. Однако представленный в статье подход делает акцент на автоматическом сборе структурированных данных и не рассматривает пользователя как основного их источника.

Данная работа построена следующим образом: в главе 2 рассказывается о примененных нами решениях проектирования и разработки и, в частности, обсуждаются преимущества и недостатки свободного ввода в стиле вики и формового ввода. Далее мы описываем процесс разработки системы, а также процесс её использования и сопровождения. В главе 3 мы показываем преимущества и возможности, которые становятся доступными благодаря использованию семантических технологий на портале. И наконец, в главе 4 приводятся результаты тестов производительности и делается сравнение Semantic MediaWiki с базовой, не семантической платформой MediaWiki. В главе 5 подводятся итоги исследования.

2 Проектирование и разработка портала

Наиболее частое применение Semantic MediaWiki и вики-систем в целом - это организация совместного управления знаниями. Хорошим примером такого применения служит сообщество semanticweb.org. В этой главе мы расскажем о портале, построенном с использованием Semantic MediaWiki, уделяя особое внимание описанию того, как при этом использовались возможности семантических технологий.

2.1 Свободное аннотирование или управляемый ввод данных?

Вики-системы предоставляют пользователям средства простого добавления и редактирования содержимого: все, что требуется от пользователя - это знание простой вики-разметки и наличие браузера. Таким образом, пользователи могут публиковать содержимое, не прибегая к помощи вебмастера. В процессе проектирования портала одной из наших главных целей было то, чтобы сотрудники института могли принимать участие в создании, развитии и поддержке содержимого портала. В связи с этим мы сочли вики правильным выбором. В отличие от обычных вики-систем Semantic MediaWiki поз-

воляет семантически аннотировать содержимое сайта. Эта свободная и независимая модель аннотирования гибка и легко расширяема. Кроме того, она не требует никакого знания определенной заранее схемы данных – эта схема может быть добавлена в вики позже. Отсюда следует важный принцип, состоящий в том что большее количество аннотаций зачастую лучше, чем малое – даже если они и не очень хорошо организованы, не следуют никакому определенному заранее словарю или онтологии. Однако при использовании *встроенных запросов* (см. главу 3.1) необходимо знать конкретные названия семантических свойств, ведь формальные запросы должны быть предельно точными, в них не допускаются даже минимальные отклонения. То же самое касается и большинства приложений, которые используют структурированные данные. Обычно они используют некоторую схему данных или словарь терминов. Следовательно, необходимо найти баланс между использованием предопределенной схемы данных и поддержанием её гибкости и расширяемости.

В Semantic MediaWiki используются *шаблоны* и *формы ввода*², что позволяет задать ограничения на используемое множество видов аннотаций. Шаблоны задают логику и внешний вид части страницы. Внутри шаблона используются переменные, значения которых подставляются при вызове шаблона на странице. Если в тексте шаблона встречаются семантические свойства, то аннотируются все страницы, использующие этот шаблон. Следовательно, при изменении значений свойств в шаблоне каскадно обновляются все зависимые от него страницы. Это очень гибкое средство для модификации структурированных данных в вики. Формы ввода предоставляют графический интерфейс для заполнения шаблонов корректными значениями, а их использование не требует даже знания вики-разметки. Таким образом, сочетание возможностей шаблонов и форм позволяет нам иметь множество определенных заранее видов аннотаций.

При разработке портала мы создали около 30 шаблонов для основных, часто повторяющихся видов ресурсов: *людей, лекций, публикаций* и т. д. На рисунке 1 показан пример формы, используемой для редактирования страниц о проектах. Формы могут содержать разные типы полей ввода: поля для ввода текста, дат, выпадающие списки и т.д. Каждому полю соответствует аннотация, то есть семантическое свойство. При вводе значения в поле, оно присваивается соответствующему свойству. Хорошей практикой является импорт этих свойств из уже существующих словарей,³ например импорт словаря FOAF⁴ для описания людей. Для того, чтобы оставить возможность для свободного, аннотирования, в формах используются поля ввода, в которых можно размещать аннотированный текст. Посредством этого мы старались найти баланс между управляемым вводом с определенными заранее семантическими свойствами и возможностью размечать текст свободно.

² http://www.mediawiki.org/wiki/Extension:Semantic_Forms

³ http://semantic-mediawiki.org/wiki/Help:Import_vocabulary

⁴ <http://xmlns.com/foaf/spec/>

The image shows a web application interface. On the left is a data entry form titled 'Projekt'. It contains several fields: 'Kurzname:' with value 'iGreen', 'Name:' with value 'iGreen', 'Name EN:' with value 'iGreen', 'Beschreibung DE:' with a text area containing German text about the iGreen project, 'Beschreibung EN:' with a text area containing English text about the iGreen project, 'Kontaktperson:' with value 'Duc Thanh Tran', 'URL:' with value 'http://www.igreen-projekt.de', and 'Start:' with a date picker set to April 2009. On the right is a rich text editor showing a sample text about Cloud Computing. The text includes a title, a date, and a paragraph of text. The editor has a toolbar at the top with various icons for text formatting and a scroll bar on the right.

Рис. 1. Слева показан ввод данных с помощью формы. Форма состоит из нескольких полей ввода. Содержимое каждого поля присваивается в качестве значения к семантической аннотации. Также в формах есть поля для свободного ввода текста, в которых может находиться текст с произвольными аннотациями. С правой стороны показан пример полностью редактирования содержимого в стиле вики, лишенного каких-либо ограничений по использованию аннотаций.

Преимуществом такого смешанного типа аннотирования является то, что структура данных может развиваться динамически, о чем подробно рассказывается в разделе 2.3.

2.2 Роли пользователей

Когда мы предложили использовать вики для создания портала, первой реакцией наших коллег был страх того, что в вики *любой пользователь* имеет права на редактирование, даже анонимные посетители. Конечно же, этот страх не был обоснован. Мы использовали систему разграничения доступа, встроенную в MediaWiki⁵ и создали четыре группы пользователей:

1. Анонимные посетители портала имеют право на чтение обычных вики-страниц, то есть статей, находящихся в основном пространстве имен.⁶
2. Зарегистрированные на портале пользователи получают возможность читать страницы в других пространствах имен. Они также обладают правами на редактирование вики-страниц, за исключением тех из них, что являлись формами или шаблонами.

⁵ http://www.mediawiki.org/wiki/Manual:User_rights

⁶ См. подробнее о пространствах имен в MediaWiki на странице <http://www.mediawiki.org/wiki/Manual:Namespaces/ru>. – Прим. перев.

3. Администраторы вики имеют также право на редактирование форм и шаблонов.
4. Наконец, четвертая группа пользователей называется бюрократами. Бюрократы имеют те же права, что и администраторы, плюс они имели право давать администраторские права и лишать их.

Необходимость в регистрации и заведении очередной учетной записи может отпугнуть пользователей, поэтому мы использовали расширение MediaWiki для аутентификации по *LDAP* (Lightweight Directory Access Protocol, облегченный протокол доступа к каталогам). Соединение между LDAP-сервером и порталом осуществлялось по зашифрованному протоколу *SSL*. Это позволяло использовать уже существующие учетные записи пользователей для аутентификации на портале.

2.3 Разработка и динамика представления структурированных знаний

Работы по разработке портала можно условно разбить на четыре различных области: настройка системы, визуальное оформление и кастомизация, импорт данных, и, наконец, разработка шаблонов, включающая в себя моделирование структурированных данных, то есть свойств и классов.

Настройка ядра Semantic MediaWiki занимает меньше часа⁷; время на разработку скина (темы оформления) зависит от решаемой задачи. В нашем случае потребовались около 80 часов времени студента-разработчика, для того, чтобы привести 148 страниц организации к единому оформлению.⁸

Чтобы подсчитать затраты на разработку мы учли правки (то есть, ревизии) страниц вики, а также созданные в ходе разработки новые страницы. На графиках 2 и 3 показана динамика изменения количества правок, где каждая точка обозначает количество правок, произведенных за месяц. Также на графиках показаны основные этапы цикла разработки портала: разработка, выпуск версии, пригодной к тестированию и, наконец, финальный релиз портала.

Динамика представления структурированных знаний Как уже говорилось в разделе 2.1, Semantic MediaWiki предоставляет средства создания состоящей из классов и свойств гибкой структурированной схемы данных портала. На графике 2 показано, как эти элементы изменялись с течением времени и то, как часто добавлялись новые элементы. В MediaWiki для группировки страниц используются категории. С точки зрения представления структуры категории соответствуют классам. Внимательно рассмотрев статистику,

⁷ http://www.mediawiki.org/wiki/Manual:FAQ#How_do_I_install_MediaWiki.3F

⁸ Базил поделился некоторыми подробностями того, как прошли эти 80 часов. Дело в том, что в университете Карлсруэ есть огромный документ под названием "Рекомендации по оформлению веб-сайтов". В этом наборе гайдлайнов указано до мелочей все параметры веб-дизайна: от оттенка бирюзового цвета на логотипе до величины отступов. — прим. перев.

касающуюся классов и свойств, можно увидеть, что большая часть работы, касающейся структуры портала, была проделана в самом начале проекта, в апреле 2009 года. В частности, предполагаемые к использованию классы были известны с самого начала и их иерархия подверглась сравнительно небольшим изменениям в течение последующих фаз разработки проекта. То же самое касается и семантических свойств. Тем не менее, можно видеть, что небольшое количество свойств и классов добавлялось и подвергалось изменениям в течение проекта. Особенный случай представляют пики в марте 2010 года. В этом месяце институт подготавливал годовой отчет о сотрудниках, публикации и прошедших событиях. Данные для отчета были экспортированы из портала. Редакторы попросили внести некоторые изменения и добавления в данные, например разделить поле "Имя" на "Фамилию и Имя". В связи с этим мы должны были соответствующим образом изменить структуру портала. В частности, структура классов подверглась переработке, такой как разделение класса *Работник* на бывших и нынешних работников.

Все эти настройки были сделаны "на ходу" в ответ на поступающие требования и запросы. В частности, следует помнить, что все изменения происходили непосредственно на уровне приложения. Мы никогда не сталкивались с ситуацией, когда прямое изменение базы данных являлось бы единственным вариантом решения проблем, также нам никогда не приходилось переводить систему в автономный режим для того, чтобы провести какие-либо модификации. Кроме того, вики предоставляет систему контроля версий, которая отслеживает и записывает все изменения на вики, в том числе изменения, касающиеся свойств и классов - а это важнейшая часть функционала семантических порталов [3].

страницы	16.716
шаблоны	219
формы	30
загруженные файлы	1.773 (1.2 GB)
пользователи (всего)	142
активные пользователи (за последние 90 дней)	83
аннотации (экземпляры свойств)	104.182
типы свойств	191
категории (классы)	40
OWL/RDF	238k triples
коддовая база	132 MB
база данных	99.5 MB

Таблица 1. Количественные характеристики портала (по состоянию на июнь 2010)



Рис. 2. На графике слева показано количество новых типов свойств и количество редакций в месяц. Правый график демонстрирует количество новых категорий и правки них. Категории соответствуют классам структурированных данных. В связи с тем, что классы и свойства являются элементами структурированных данных, данные графики показывают эволюцию данных в течение времени.

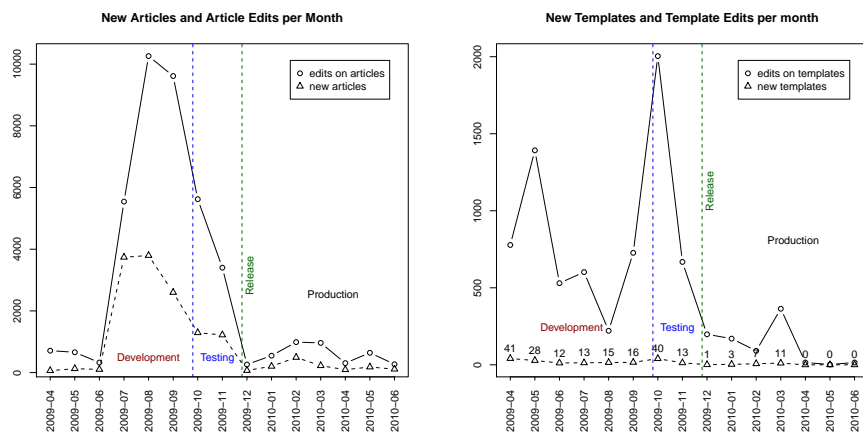


Рис. 3. На левом графике показано количество новых статей и редакций в статьях в месяц и показаны периоды развития портала - от разработки до выпуска релиза. В связи с тем, что период разработки включал в себя автоматическое наполнение портала содержимым, на графике имеется пик, приходящийся на эту стадию. График справа демонстрирует зависимости появления новых шаблонов и редакций существующих от времени. Так как право редактирования шаблонов есть только у администраторов, по графику можно оценить затраты на разработку портала и поддержку его после релиза. Пик в марте 2010 - результат работы над ежегодным отчетом, см. главу 2.3.

допустимы, однако к официальному представительству института в WWW предъявляются намного более жесткие требования. К примеру, в шаблонах все переменные, оставшиеся незаполненными, должны быть скрыты. Кроме того, в связи с высокими требованиями к визуальному оформлению, некоторые аннотации содержали вики-разметку (например курсивный шрифт или указания на размер кегля), что существенно помогало представлению структурированных данных. Более того, в связи с тем, что шаблоны являются одновременно и инструментом визуального оформления вики-статьи, и содержат логику отображения, манипуляции с шаблонами становятся довольно сложным занятием, требующим высокого уровня владения вики-разметкой. В связи с этим, манипуляции с шаблонами были запрещены администраторами портала.

3 Где может помочь семантика? Возможности Semantic MediaWiki

В предыдущей главе был описан процесс разработки и динамика структурированных данных. В этой главе мы описываем функциональные возможности, использующие их преимущества.

3.1 Встроенные запросы

Наибольшим преимуществом SMW, помимо гибкой парадигмы аннотаций, является возможность повторного использования данных внутри платформы. Это достигается путем использования запросов, собирающих информацию со страниц вики. Эти *встроенные запросы* позволяют извлекать множества данных или отдельные значения семантических свойств, предоставляя возможность их визуального представления посредством таблиц, списков, карт и т.д. Такое повторное использование данных позволяет избежать их дублирования. К примеру, информация о человеке (его имя, адрес электронной почты или номер телефона и т.д.) вводится единожды. Когда эта информация требуется на других вики-страницах – проектах этого человека, его публикациях и т.д. – она автоматически запрашивается и выводится в требуемом формате. При изменении данных на странице-источнике, происходит соответствующее им обновление на всех страницах, запрашивающих эти данные. Фактически, встроенные запросы позволяют создавать динамические страницы. На рисунке 4 представлен пример встроенного запроса и результаты его выполнения.

3.2 Запросы к ресурсам Связанных Открытых Данных (Linked Open Data)

Мы создали расширение для MediaWiki, которое позволяет извлекать данные из внешних источников, используя простой синтаксис встроенных запросов[7].

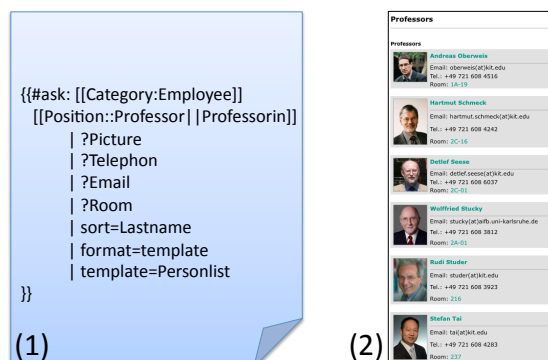


Рис. 4. Внешний вид запроса всех сотрудников, которые работают профессорами с выводом информации о них (1) и результаты этого запроса(2).

Это расширение выполняет роль посредника между порталом и различными внешними источниками данных – базой знаний Freebase, другими SMW, файлами в формате CSV. Оно позволяет импортировать данные, представленные в этих источниках, обогащая таким образом содержимое портала. В случае с Freebase, программа-посредник транслирует встроенный запрос SMW в MQL - язык запросов к Freebase. На рисунке 5 показан пример такой трансляции. Важно отметить, что задача трансляции не имеет чисто лексический характер, а включает в себя отображение онтологий (ontology mapping). Результаты отображения хранятся в вики в виде аннотаций. Таким образом в их составлении и сопровождении могут участвовать пользователи системы.

На нашем портале мы также осуществляли запросы к SMW semanticweb.org для того, чтобы получать последние новости о научных событиях, таких как конференции и семинары. Мы представляли эти данные в виде линии времени, с нанесенными на неё событиями. Пользователи, таким образом, получают интерактивный, предоставляющий актуальную информацию инструмент для отслеживания конференций. Кроме того, мы использовали Freebase для получения геолокационной информации о промышленных и академических партнерах института, что позволяло осуществлять их сортировку по региону.

3.3 Использование семантической информации при поиске

Несомненным преимуществом доступности на портале структурированных данных является возможность использовать их при поиске. Для организации семантического поиска на портале¹¹ мы использовали подход, описанный в [8]. Данный подход позволяет использовать при поиске ключе-

¹¹ <http://www.aifb.kit.edu/web/Spezial:ATWSpecialSearch>

i)	<pre>{ask: [[Category:Country]] [[located in::Europe]] ?has capital source=freebase }}</pre>	ii)	<pre>{ "limit" : 50, "type/object/name" : null, "location/country/capital" : [{ "name" : null }], "type/object/type" : "location/country", "location/location/containedby" : [{ "type/object/name" : "Europe" }] }</pre>
iii)	<pre>Category country → /location/country Property has capital → /location/country/capital Property located in → /location/location/containedby</pre>		

Рис. 5. Использование программы-посредника для общения с базой знаний Freebase. Встроенный запрос i) преобразуется к MQL-запросу ii), используя информацию об отображении iii).

вые слова, что привычно для многих пользователей. Ключевые слова затем преобразуются в интерпретации - при этом в качестве пространства поиска используется множество структурированных данных вики. Возможные интерпретации показываются пользователю; он же, в свою очередь, выбирает ту из них, которая в наибольшей степени отвечает поисковому запросу. Последний шаг состоит в уточнении результатов запроса. На рисунке 6 показан пример поиска по структурированным данным. В качестве объектов поиска выступают работники, адреса их электронной почты и расположение их рабочих мест. Встроенные запросы имеют простой но формальный синтаксис, что делает их неприменимыми для свободного поиска. В отличие от них, разработанный нами инструмент *Спроси у вики*, с одной стороны, подходит для нужд конечных пользователей, а с другой - использует все преимущества семантических аннотаций.

4 Производительность

MediaWiki заслужила репутацию хорошо масштабируемой и быстрой платформы - её использует Википедия, она применяется как основа многих других сайтов. Возможности, предоставляемые Semantic MediaWiki, привлекают внимание широкого круга потенциальных пользователей, однако многие из них скептически настроены в вопросах, связанным с потреблением ресурсов, стабильностью и масштабируемостью платформы. В этой главе мы описываем результаты нагрузочного тестирования, произведенные как на Semantic MediaWiki, так и на MediaWiki с использованием информации, представленной на портале. Благодаря таким тестам мы получаем возможность сравнения этих двух платформ.

Условия проведения испытаний Тесты производились с использованием стандартного настольного компьютера с частотой процессора 2ГГц, с двумя гигабайтами памяти и установленной ОС Debian 5.0¹². Вики была запущена

¹² AMD Athlon 64 3200+, 2.6.26-2-amd64 kernel

Ask The Wiki

New Search | Step 1: Enter keywords → Step 2: Choose interpretation → **Step 3: View and refine results**

- There are **108** results matching your interpretation.
- Use the **Facets** to the right to expand or narrow the results.

Legend: Concepts, Relations, Labels, Literals

Thomas Karle	thomas.karle(at)aifb.uni-karlsruhe.de	1A-13	05.20
Alaa Ismaeel	ais(at)aifb.uni-karlsruhe.de	2B-09	05.20
Duc Thanh Tran	ducthanh.tran@kit.edu	260	11.40

Facets

- Open/close a menu.
- Remove a concept.
- Add/remove a relation.
- Define a concept.

Mitarbeiter

- workplaceHomepage
- sameAs
- depiction
- Land
- Public Key
- Stellung
- Info
- mbox
- Bundesland
- lastName
- Mobile
- Telefax

Рис. 6. На рисунках показаны результаты поиска работников, их адресов электронной почты, номеров их кабинетов и соответствующих номеров зданий. Меню фильтров справа позволяет уточнить результаты поиска на основе структурированных данных.

на веб-сервере Apache web с PHP 5 использовала СУБД MySQL¹³. Нагрузочные тесты управлялись с помощью Apache JMeter, на стороне сервера была запущена программа мониторинга sysstat¹⁴. Клиент посылал запросы будучи соединенным с той же стомегабитной магистралью, к которой был подключен сервер.

Такая системная конфигурация является типичной для установки системы на одну машину и никак не может использоваться в реальных системах с высокими требованиями по производительности. Тем не менее, она позволяет сравнить две системы, MediaWiki 1.15.3 и Semantic MediaWiki 1.5. Вики хранит данные о портале AIFB см. Табл. 1. Эти данные содержат аннотации, представленные в синтаксисе SMW. Если расширение SMW не подключено, то MediaWiki интерпретирует эти инструкции как простой текст. SMW позволяет задать ограничения на использование встроенных запросов, например максимальное число условий поиска, глубину запроса и максимальную величину выбираемой из БД порции данных. Эти ограничения были взяты как метрики производительности. Для измерений были выбраны 310 статей (около 2% от общего числа страниц), доступных с главной страницы портала в один-два клика. Данное подмножество страниц является репрезентативной выборкой из генеральной совокупности всех страниц портала, ибо включает в себя как страницы с небольшим количеством семантических аннотаций и запросов, так и статьи, активно использующие эти возможности. В среднем страница содержала 10 встроенных запросов и 12 семантических аннотаций.

¹³ Apache 2.2.9, PHP 5.2.6-1+lenny3 с APC 3.0.19, MySQL 5.0.32

¹⁴ Apache JMeter 2.3.4, sysstat 7.0

4.1 MediaWiki против Semantic MediaWiki

В процессе тестирования N пользователей параллельно запрашивали 310 страниц в случайном порядке. На рисунке 7 представлены графики времени отклика системы в миллисекундах для $N = \{1, 10, 25, 50\}$ параллельно работающих пользователей. Графики иллюстрируют поведение MediaWiki (MW), Semantic MediaWiki (SMW), и Semantic MediaWiki с кэшированием (SMW+C). Данный график показывает, что время отклика линейно зависит от количества параллельно работающих пользователей. Это поведение особенно хорошо видно в таблице 4.1, показывающей производительность системы, т.е. количество обработанных запросов в секунду. Пропускная способность является константой с примерным значением в 4.7 запросов/сек для MW и 4.1 запросов/сек для SMW, что означает, что использование SMW приводит к примерно тринадцатипроцентному уменьшению производительности во время использования по сравнению с MW.

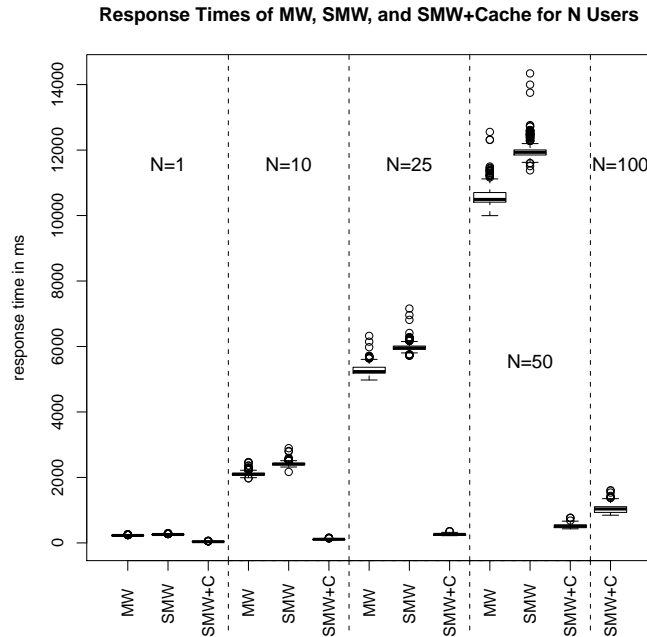


Рис. 7. График, иллюстрирующий время отклика на страницу для MW, SMW, SMW+Cache с N параллельно работающими пользователями, запрашивающих 310 страниц в произвольном порядке.

Довольно неожиданным фактом было то, что разброс значений времени отклика для 310 страниц оказался таким небольшим (см. как показано на графике 7. Особенно удивляет такая небольшая дисперсия для SMW,

где страницы содержат семантические аннотации и встроенные запросы, которые должны быть подвергнуты синтаксическому анализу и интерпретации. Время отклика на таких страницах должно зависеть от количества и сложности этих запросов. Было выяснено, что малый разброс объясняется неявным кэшированием. Веб-сервер имеет встроенные кэш PHP кода (APC), который используется MW, а следовательно и SMW. Также следует заметить, что содержимое страницы не подвергается интерпретации при каждом запросе, этот процесс происходит только при необходимости. База данных также кэширует SQL-запросы с использованием подсистемы InnoDB. Все эти встроенные средства кэширования берут на себя большую часть накладных расходов SMW во время функционирования системы и делают возможным запуск SMW с константной ценой по производительности - около 13% уменьшения производительности по сравнению с несемантической MediaWiki, см. Рис. 4.1. Тестирование показало, что ресурсы процессора являются узким местом как для MW так и для SMW для всех запусков, в которые было вовлечено больше одного пользователя. Ресурсы процессора потреблялись на 95% веб-сервером и 5% базой данных.

N	1	10	25	50	100
MW	4.36	4.75	4.75	4.73	n/a
SMW	3.83 (-12 %)	4.10 (-14%)	4.13 (-13 %)	4.13 (-13%)	n/a
SMW+C	> 25.68 (+489%)	> 90.80 (+1810%)	> 96.78 (+1930%)	> 96.31 (+1930%)	> 95.01

Таблица 2. Производительность (запрос/сек) для N параллельно работающих пользователей. Проценты указаны в сравнении с производительностью MediaWiki (MW). При использовании кэширования (SMW+C) пределы производительности сервера не были достигнуты.

Для того, чтобы избежать неявного кэширования и провести измерения истинных значений потребления ресурсов, мы произвели также *холодный* тест: мы перезагружали машину после того как каждая страница была запрошена и повторили этот тест десять раз. На рисунке 8 показано среднее время отклика для десяти запусков для каждой страницы. Страницы отсортированы по возрастанию количества использованных встроенных запросов и по количеству шаблонов. Можно видеть, что для MW время отклика немного увеличивается при возрастании количества шаблонов на страницу. В случае с SMW можно сказать, что в общем большее количество запросов на страницу увеличивает время отклика. Следует заметить, однако, что время отклика зависит в наибольшей степени от вида конкретного запроса, что ясно демонстрируют высокие пики на графике. Наибольшее значение наблюдалось при обработке страницы, содержащий запрос списка всех людей. Данная операция требовала запроса фотографий для каждого сотрудника с последующим созданием её миниатюры. Аналогичная ситуация наблюдает-

ся и в случае с остальными пиками - соответствующие страницы содержат по одному встроенному запросу, связанному с обработкой изображений. Запросы, результатами которых является текстовая информация, обходятся намного дешевле (в смысле быстродействия). Так например для страниц, содержащих 20 и более встроенных запросов их обработка занимает менее двух секунд, если не требуется извлечение графических данных. Однако изображения являются статическим типом контента и легко кэшируются. Мы обсуждаем применённые техники кэширования ниже.

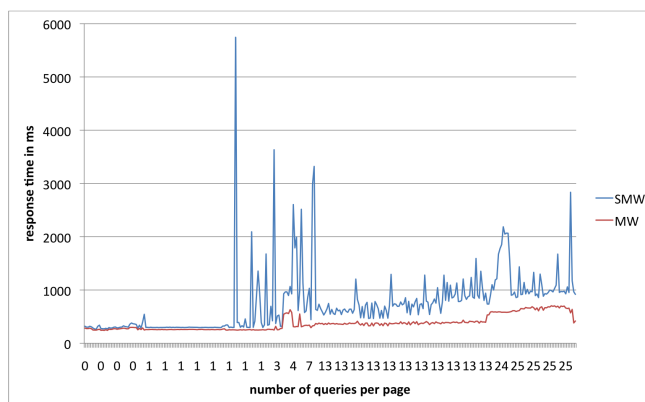


Рис. 8. Время отклика при холодном тесте для 310 страниц, отсортированных по возрастанию количества использованных встроенных запросов. Большие всплески в центре графика созданы встроенными запросами, использующими работу с изображениями

4.2 Кэширование динамических страниц

Для того чтобы ускорить быстродействие веб-сайта и уменьшить нагрузку на веб сервер были применены обратные прокси-сервера (reverse proxy). Обратный прокси-сервер - это программа, установленная перед веб-сервером и кэширующая данные. Его функция состоит в том, чтобы возвращать запрошенное содержимое, если оно доступно в кэше и перенаправлять запрос к веб-серверу в противном случае. Популярный веб-кэшем является Squid¹⁵, он поддерживается движком MediaWiki, как схематически изображено на рис. 9. Эта система хорошо работает для статического контента, например документов HTML и графических файлов, однако дело усложняется, если требуется кэширование динамического содержимого. В контексте Semantic MediaWiki основными источниками динамического содержимого

¹⁵ <http://www.squid-cache.org>

являются встроенные запросы к страницам вики или внешним ресурсам через службы-посредники. Страницы, содержащие семантические запросы являются динамическими в том смысле, что содержимое страницы изменяется с течением времени, в то время как исходный код страницы остается неизменным. Следовательно возникает проблема, состоящая в том, что поле **Last-Modified** в HTTP-заголовке остаётся неизменной, т.к. веб-сервер не распознаёт изменения. Эта запись используется системой кэширования для того, чтобы определить, является ли обозреваемая страница свежей, а потому нам было необходимо изменить механизм кэша. Мы выбрали агрессивную стратегию кэширования: не обращая внимание на поле **Last-Modified**, мы установили максимальное время устаревания страницы в три часа. Таким образом невидимые изменения динамических страниц отображались максимум через этот период. Если же страница редактировалась напрямую, она немедленно удалялась из кэша. Изображения и другое статическое содержимое кэшируется на более длительные периоды. Применение кэша влечет огромный прирост по производительности - около 90 запросов/сек, как показано в Таблице 4.1 Рис. 7. Однако, это далеко не является пределом, поскольку процессор при тестах использовался лишь на 30%, даже для ста параллельно работающих пользователей. Для того, чтобы измерить предел производительности системы с кэшированием, потребовалось бы подключить к ней несколько физических клиентов, посылающих запросы.

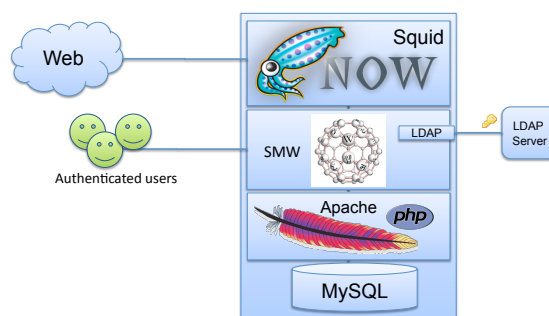


Рис. 9. Инфраструктура портала. Анонимные пользователи получают содержимое из кэша Squid при его доступности. Аутентифицированные пользователи подключены к веб-серверу напрямую.

4.3 Производительность в действии

Портал работает в режиме онлайн в течение более чем шести последних месяцев и единственным сбой в сего работе был вызван лишь единожды DoS-атакой. Проблема была решена ограничением количества соединений с веб-сервером. Мы заключаем из этого, что предложенное нами решение

является стабильным и довольно робастным. Количество посещений в день колеблется между 60 и 120 тысячами в день, что создает шестипроцентную нагрузку на процессор и среднюю загрузку системы равную 0.1. Веб-сервер запущен на компьютере с двумя процессорами¹⁶ с двумя гигабайтами памяти. Среднее значение времени отклика (в этом случае мы имеем в виду время между приходом запроса и выдачей ответа сервером) приблизительно равно двум мс. Запрос данных, не находящихся в кэше занимает в среднем около 200 мс. Запросы, обращенные к кэшу составляют 80% от всего множества запросов, а размер кэша на диске равен приблизительно 550MB.

5 Заключение

В этой статье мы показали способ применения парадигмы совместного редактирования содержимого в вики к созданию веб-портала, использующего семантические технологии. В частности мы показали как можно одновременно сочетать техники свободного семантического аннотирования и использование заранее определенных аннотаций для достижения гибкости и расширяемости структурированных данных. Далее мы показали то, как структурированные данные, доступные из аннотаций развиваются со временем, продемонстрировав то, как возможно изменять и расширять их в процессе работы, не затрагивая базу данных. Использование семантических данных и польза от функциональности Semantic MediaWiki иллюстрируется несколькими примерами. Наконец мы оценили производительность системы и сравнили её с несемантической альтернативой, показав при этом, как кэширование может резко поднять производительность. Принимая во внимание все вышесказанное можно видеть, что Semantic MediaWiki может быть использована в качестве мощной платформы для создания порталов, успешно применяющих семантические технологии.

6 Благодарности

Мы выражаем особую благодарность Мартину Джангу чья преданность проекту составила значительную часть его успеха. Спасибо Николь Арльт и Фабио Гарзотто за их вклад, а также Филиппу Зоргу и IT-подразделению AIFB за их ценные замечания, отзывы и техническую поддержку. Представленная в этой статье работа поддержана грантом EU IST FP7 проекта ACTIVE. Номер гранта 215040.

7 Перевод

Перевод на русский язык выполнен Катковым Юрием.

¹⁶ Intel(R) Xeon(R) CPU E5450 @ 3.00GHz

Список литературы

1. Maedche, A., Staab, S., Stojanovic, N., Studer, R., Sure, Y.: Semantic portal - the seal approach. In Fensel, D., Hendler, J., Lieberman, H., Wahlster, W., eds.: *Spinning the Semantic Web*. MIT Press, Cambridge, MA. (2003) 317–359
2. Hotho, A., Maedche, A., Staab, S., Studer, R.: Seal-II - the soft spot between richly structured and unstructured knowledge. *Journal of Universal Computer Science* **7**(7) (2001) 566–590
3. Lara, R., Han, S.K., Lausen, H., Stollberg, M., Ding, Y., Fensel, D.: An evaluation of semantic web portals. In: *IADIS Applied Computing International Conference*. (2004) 23 – 26
4. Corlosquet, S., Delbru, R., Clark, T., Polleres, A., Decker, S.: Produce and consume linked data with drupal! In: *8th International Semantic Web Conference (ISWC2009)*. Volume 5823 of LNCS., Springer (2009)
5. Krötzsch, M., Vrandečić, D., Völkel, M., Haller, H., Studer, R.: Semantic wikipedia. *Journal of Web Semantics* **5**(4) (2007) 251–261
6. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic mediawiki. In: *Proceedings of the 5th International Semantic Web Conference (ISWC2006)*. Volume 4273 of LNCS., Springer (2006) 935–942
7. Ell, B.: Integration of external data in semantic wikis. Master’s thesis, Hochschule Mannheim (2009)
8. Haase, P., Herzig, D.M., Musen, M., Tran, D.T.: Semantic wiki search. In: *6th European Semantic Web Conference (ESWC2009)*. Volume 5554 of LNCS., Springer Verlag (2009) 445–460